Densidées : calcul automatique de la densité des idées dans un corpus oral

Hyeran Lee[1], Philippe Gambette[2], Elsa Maillé [3], Constance Thuillier[3]

(1) Praxiling - Université Montpellier 3 / CNRS, 17 rue Abbé de l'Épée 34090 Montpellier France
(2) LIRMM - Université Montpellier 2 / CNRS, 34095 Montpellier Cedex 5
(3) ISTR - Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex
hlee1@univ-montp3.fr

Résumé La densité des idées, qui correspond au ratio entre le nombre de propositions sémantiques et le nombre de mots dans un texte reflète la qualité informative des propositions langagières d'un texte. L'apparition de la maladie d'Alzheimer a été reliée à une dégradation de la densité des idées, ce qui explique l'intérêt pour un calcul automatique de cette mesure. Nous proposons une méthode basée sur un étiquetage morphosyntaxique et des règles d'ajustement, inspirée du logiciel CPIDR. Cette méthode a été validée sur un corpus de quarante entretiens oraux transcrits et obtient de meilleurs résultats pour le français que CPIDR pour l'anglais. Elle est implémentée dans le logiciel libre *Densidées* disponible sur http://code.google.com/p/densidees.

Abstract Idea density, which is the ratio of semantic propositions divided by the number of words in a text, reflects the informative quality of the sentences of a text. A decreasing idea density has been identified as one of the symptoms of Alzheimer's disease, which explains the interest in an automatic calculation of idea density. We propose a method based on part-of-speech tagging followed by adjustment rules inspired from the CPIDR software. This method was validated on a corpus of 40 transcribed conversations in French and obtains better results in French than CPIDR in English. It is implemented in the free software *Densidées* available at http://code.google.com/p/densidees.

Mots-clés : densité des idées, analyse prédicative, étiquetage sémantique, psycholinguistique
Keywords: idea density, propositional analysis, semantic tagging, psycholinguistics

1 Introduction

La dégradation de la capacité linguistique est un élément caractéristique de la maladie d'Alzheimer (Moreaud et *al.*, 2001). Le déficit linguistique dans cette pathologie est dû à l'altération de la mémoire sémantique. Ainsi, pour évaluer la performance langagière des patients atteints de la maladie d'Alzheimer, une mesure fine au niveau sémantique du langage est nécessaire. La densité des idées (DI), ratio pour dix mots du nombre de propositions sémantiques dans un texte, est aujourd'hui reconnue comme un indicateur pertinent des fonctions intellectuelles de sujets. Cette méthode a été validée par de nombreuses études psycholinguistiques appliquées : la compréhension du texte (Kintsch et *al.*, 1973 ; Kintsch, 1978), la mémoire (Thorson et *al.*, 1984), la maladie d'Alzheimer (Snowdon et *al.*, 1996 ; Lee et *al.*, 2009), la qualité de prise de note des étudiants (Takao et *al.*, 2002), le vieillissement (Kemper et *al.*, 2001), la schizophrénie (Covington et *al.*, 2007), le genre du discours (Covington, 2009). Cependant, son analyse longue, fastidieuse, et parfois subjective fait souvent obstacle à son utilisation pratique. Par conséquent, il serait intéressant et innovant de développer un outil automatique permettant de donner un résultat rapide et fiable de la DI. Nous proposerons une méthode de calcul automatique de la DI, basée sur un étiquetage morphosyntaxique et des règles d'ajustement, inspirée du logiciel CPIDR (Brown et *al.*, 2008). Enfin nous testerons la validité de cette méthode sur un corpus de quarante entretiens oraux transcrits.

1.1 Analyse prédicative et densité des idées

Pour étudier les mécanismes qui sous-tendent la pensée humaine, la psychologie cognitive n'ayant accès qu'aux productions du sujet, considère qu'elles possèdent en leur sein les marques des mécanismes qui les ont engendrées. En ce sens, les productions discursives sont privilégiées car le langage est à la fois le support et le produit de la pensée. Autrement dit, les structures langagières peuvent refléter les structures cognitives. Ainsi, une certaine forme d'analyse du discours revient à modéliser des phénomènes cognitifs.

La fonction primitive du langage est la fonction référentielle (Jakobson, 1963), c'est-à-dire qu'il sert à transmettre à autrui des informations du monde réel par la symbolisation. Cette fonction est accomplie en véhiculant du sens. L'activité sémantique consiste donc à produire du sens dans l'intellect du récepteur, c'est-à-dire la formation de représentation mentale chez l'interlocuteur par l'intermédiaire du langage. Depuis la logique aristotélicienne ainsi que dans la logique classique de Frege (1967, 1971), en passant par la théorie psychologique des réseaux sémantiques propositionnels d'Anderson (1976), les chercheurs se sont intéressés au traitement sémantique de l'information et ont tenté de définir la structure cognitive. Ils ont fait l'hypothèse que l'information dans la mémoire est organisée sous forme propositionnelle. En effet, un mot isolé seul ne suffit pas à créer une idée, c'est l'ensemble de propriétés et de relations s'y rapportant qui permet d'appréhender et de produire la signification psychologique. À partir de ces théories logiques et psychologiques, Kintsch (1974) a développé une méthode d'analyse linguistique qui permet de modéliser la manière dont l'humain encode les informations, appelée *analyse prédicative*. Il part du postulat que la forme dominante de la représentation cognitive du langage est de nature propositionnelle. Ainsi, « *si l'on considère que la prédication qui s'exprime dans un message linguistique est une activité cognitive essentielle de l'homme et que, sous-jacent à la réalisation de surface, c'est-à-dire au mot, se trouve un concept, on peut estimer que l'analyse prédicative, outil de description sémantique des textes, est pour le psychologue la transcription d'une activité cognitive* » (Ghiglione et *al.*, 1995 : 49).

L'analyse prédicative permet d'extraire les propositions sémantiques dans le discours par la concaténation des unités élémentaires du sens : prédicat et argument(s). Par exemple, dans la phrase « Le chien poursuivait un chat dans le jardin », exemple emprunté à Le Ny (1989), les concepts génériques qui font référence à des objets (« chien », « chat », « jardin »), à des événements (« poursuivre »), et à des relations

dans l'espace (« dans ») peuvent être extraits. On parle des *arguments* qui sont des entités référentielles pouvant correspondre à des êtres ou des objets, et des *prédicats* qui sont des unités requérant des arguments. Ainsi les prédicats assignent des propriétés aux arguments ou définissent la relation entre les arguments (Coirier et *al.*, 1996). L'analyse prédicative de cette phrase peut être notée selon la forme classique ci-dessous :

P1. POURSUIVRE (a1, a2) a1= chien, a2= chat
P2. DANS (P1, a3) a3= jardin
L'ensemble constitué d'un prédicat et de son ou ses arguments forme une *idée* ou *proposition sémantique*.

Si l'analyse prédicative reflète l'activité cognitive, la densité des idées, quant à elle, permet de la quantifier. En ramenant le nombre de propositions sémantiques au nombre de mots produits dans le discours (multiplié par 10 pour obtenir une DI pour 10 mots), on peut mesurer la densité des idées d'un discours. La DI permet donc de mesurer la quantité informative dans le discours. Une DI élevée peut refléter l'aptitude d'un locuteur à exprimer efficacement ses idées ainsi que leur interrelation complexe. Par contre, une faible DI dans le discours peut révéler un discours peu efficient, du fait de l'utilisation d'un plus grand nombre de mots pour exprimer les idées essentielles. Le score de DI de la phrase « Le chien poursuivait un chat dans le jardin » précédemment évoquée est donc de 2.5 (2/8*10 = 2.5 : 2 propositions sémantiques divisées par 8 mots, multiplié par 10).

$$\mathbf{DI} = \frac{\textbf{nombre de propositions}}{\textbf{nombre total de mots}} \;\mathbf{x\ 10}$$

1.2 Calcul automatique de la densité des idées : *Densidées*

Les études portant sur l'analyse prédicative en langue française ont été développées principalement par Le Ny (1979), Ghiglione (1982), Denhière (1983). Cependant, il n'y a pas de méthode clairement établie pour l'analyse prédicative du français et de texte oral. Aussi, selon les auteurs, quelques divergences peuvent être observées selon l'évolution de la théorie linguistique, par exemple la proposition de Le Ny (1987) d'intégrer les acquis de la grammaire des cas de Fillmore (1968) dans l'analyse prédicative. Nous avons ainsi établi des règles d'analyse prédicative, basées sur les études précédentes (Kintsch, 1974 ; Turner et *al.*, 1977 ; Le Ny, 1979 ; Ghiglione et *al.*, 1995 ; Kemper et *al.*, 2001 ; Chand et *al.*, 2010). L'exemple ci-dessous montre notre méthode d'analyse prédicative :

> *le plus beau jour de ma vie bon bah réfléchissons c'est le jour de mon mariage voilà il y a 52 ans bientôt donc voilà bien que ça a été un mariage tout à fait simple parce que je ne je n'avais plus mes parents donc quand on s'est mariés on était 12 personnes donc vous voyez*

P1. BEAU (a1) a1= jour

P2. LE PLUS (P1)

P3. DE (P1, a2) a2= vie

P4. MON (a2)

P5. COPULE (P3, a=1)

P6. DE (P5, a3) a3= mariage

P7. MON (a3)

P8. IL Y A (a4) a4= ans

P9. 52 (a4)

P10. BIENTÔT (P8)

P11. COPULE (P7, a3)

P12. SIMPLE (a3)

P13. TOUT A FAIT (P12)

P14. BIEN QUE (P6, P11)

P15. POSSEDER (a5, a6) a5= je, a6= parents

P16. MON (a6)

P17. NE PLUS (P15)

P18. PARCE QUE (P12, P17)

P19. SE MARIER (a7) a7= on

P20. COPULE (a8, a9) a8= on, a9= personnes

P21. 12 (a9)

P22. QUAND (P19, P20)

P23. DONC (P17, P22)

23/57*10= 4.04

Les prédicats peuvent être classés en trois grandes catégories : prédicateur, modificateur, et connecteur. Les prédicateurs expriment l'action ou l'état (e.g. verbes « se marier », « être » dans notre corpus) ; les modificateurs spécifient la qualité ou la quantité de l'argument (e.g. adjectif « 52 », adverbe « bientôt ») ; les connecteurs relient les différentes idées (e.g. préposition « de », conjonction « donc », causalité « parce que », concession « bien que »). Les arguments renvoient aux objets et/ou personnes (e.g. « je », « on », « mariage »), et relèvent d'une catégorie lexicale qui a pour fonction principale la désignation d'objets, en l'occurrence les substantifs. Ainsi, une proposition peut être l'argument d'une autre proposition. Si les prédicats et les arguments peuvent être relevés facilement, il est difficile d'établir leurs relations, l'utilisation des anaphores, l'ambiguïté sémantique créée par l'emploi du pronom « on » (« on » de a7 dans P19 a comme référent visé « nous » : locutrice + son mari alors que « on » de a8 dans P20 a comme référent « les gens » incluant a7) requièrent une attention soutenue au sens véhiculé.

Plusieurs problématiques de traitement automatique des langues naturelles (e.g. traduction automatique, extraction d'informations, etc.) ont fait appel à l'analyse prédicative, avec l'objectif de représenter un texte en langue naturelle par une formule logique, en langage des prédicats du premier ordre. Divers formalismes sont par exemple présentés par François (1991). Toutefois, cette approche s'est heurtée aux limites du formalisme de représentation, et de telles analyses prédicatives d'étiquetage sémantique sont actuellement utilisées en pratique uniquement sur des tâches très spécifiques et des corpus ciblés, comme dans le système présenté par Meurs et *al.* (1998). Sur des corpus plus généraux, la couverture des bases de données sémantiques (comme *FrameNet*) est trop faible, et l'analyse sémantique conduit à des taux d'erreurs importants. Ces erreurs sont amplifiées sur les corpus oraux du type de ceux qui nous intéressent dans le contexte du calcul de la DI, du fait des contraintes qu'ils induisent. En effet, de nombreuses utilisations de l'anaphore, les phénomènes oraux particuliers tels que les mots fragmentés, les énoncés inachevés, les ratages, les reformulations, les répétitions, les interjections, les habitudes du langage (gimmicks), les pauses remplies, etc. rendent l'analyse de l'oral complexe.

Cependant, le calcul de la DI ne nécessite en fait pas de calculer l'ensemble des prédicats et de leurs arguments, mais seulement de compter les prédicats. Nous avons donc choisi d'éviter d'utiliser une approche sémantique, et d'utiliser plutôt les travaux de Brown et al. (2008) sur la langue anglaise. Ceci dans le but de concevoir une approche du comptage des prédicats par un ensemble de règles appliquées après un étiquetage morphosyntaxique du texte. Pour calculer la densité des idées, Brown et *al.* proposent le logiciel CPIDR qui étiquette chaque mot du texte comme prédicat, ou bien comme non-prédicat. L'idée principale de l'étiquetage est qu'un prédicat correspond typiquement à un verbe (prédicateur), à un adjectif, à un adverbe (modificateurs), à une préposition, ou à une conjonction (connecteurs). Ainsi, l'étiquetage morphosyntaxique est à la base du calcul approximatif de la densité des idées. Cette étape d'étiquetage morphosyntaxique, traitée dans le cas de CPIDR par le logiciel MontyLingua (Liu, 2004), est suivie d'un post-traitement à base de règles destinées à corriger les erreurs d'étiquetage morphosyntaxique qui ont une influence sur le nombre de prédicats, à traiter le cas spécifique des corpus oraux (avec une gestion basique de certaines répétitions ou auto-corrections), et enfin à ajuster le calcul du nombre de prédicats. Cette méthode est efficace en anglais, puisque CPIDR obtient généralement un meilleur accord avec un ensemble d'étiqueteurs humains que les étiqueteurs humains entre eux. Nous avons donc choisi de suivre les mêmes principes, en apportant une attention particulière au caractère oral de notre corpus, important à la fois pour nos objectifs d'utilisation de la densité des idées, même si l'outil que nous proposons est aussi destiné à l'écrit. L'implémentation de ces principes pour le français nous a fait recourir à TreeTagger (Schmid, 1994) pour l'étiquetage morphosyntaxique du texte. C'est ensuite un ensemble de 35 règles d'ajustement que nous proposons pour déterminer si un mot est un prédicat ou non. Des exemples de règles sont fournis en figure 1, elles sont intégralement décrites dans le manuel d'utilisation de *Densidées*. Dans la mesure du possible, les numéros de règles utilisés dans CPIDR ont été conservés dans *Densidées*. En outre, le logiciel *Densidées* est écrit en Python de façon commentée et très lisible. Il fonctionne en

ligne de commande mais peut également être appelé depuis Windows par l'intermédiaire d'une interface graphique montrée en figure 2.

> Règle 208 - Comparatif : "que" n'est pas proposition après "autant", "moins", "pire", "plus"
>
> Règle 301 - Verbes de liaison ("apparaître", "être", "sembler", "devenir", "paraître", "rester", "demeurer") non propositions si suivis d'un adjectif ou d'un adverbe

Figure 1 : Exemples de règles de *Densidées*

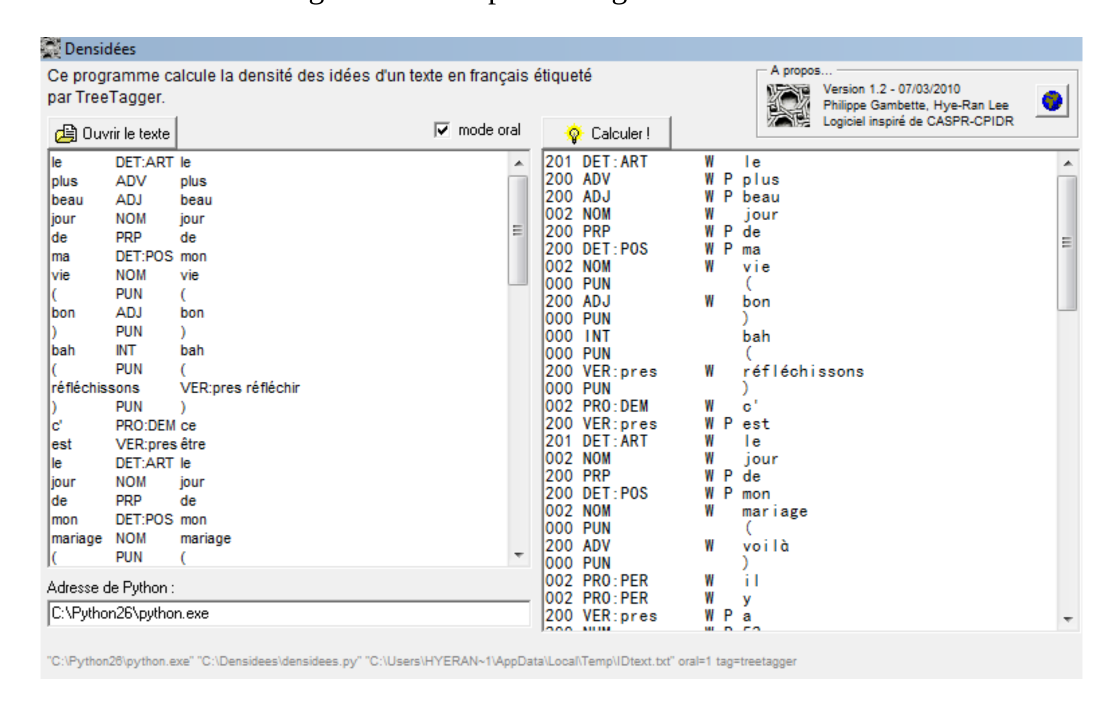


Figure 2. L'interface graphique de *Densidées* sous Windows

2 Méthode

2.1 Corpus

Pour examiner la validité de notre méthode, nous avons analysé 40 textes oraux. Ce corpus est issu d'une étude sur la description fine de la dégradation linguistique dans la maladie d'Alzheimer. Pour le recueillir, nous avons mené un entretien individuel semi-dirigé auprès de 40 sujets volontaires. Ces sujets sont âgés de 65 à 85 ans, ne présentant pas d'une pathologie cognitive, et sont locuteurs français natifs. Cet entretien dure environ 35 minutes dont 5 à 15 minutes pour l'enregistrement du discours oral. Une narration libre de l'évocation d'un souvenir personnel a été demandée pour la production du discours spontané, et une description d'image « voleur du biscuit », tirée de Boston Diagnostic Aphasia Examination (Goodglass et *al.*, 1983), pour un discours descriptif. Tous les entretiens ont été enregistrés numériquement. Ces discours oraux ont été transcrits individuellement avec une transcription orthographique standard de type GARS (Blanche-Benveniste, 1998), c'est-à-dire sans renormalisation de la parole (e.g. sans introduire de l'élément absent « ne » dans « il y a pas »), aligner la parole et le texte.

Ce corpus a été tronqué de manière à garder environ 300 mots par transcription pour que les corpus soient comparables statistiquement. Kemper et *al.* (2001) recommande que l'échantillon du discours ne soit pas trop bref pour avoir un résultat fiable, et que l'analyse porte sur un minimum de 10 énoncés, ce qui bien est le cas ici.

2.2 Procédure

Nous avons choisi de proposer un prétraitement manuel du corpus pour marquer certaines caractéristiques spécifiques à l'oral qui ne semblent difficilement traitables de façon automatique. Nous avons utilisé les crochets [ ] pour marquer les mots fragmentés, les répétitions successives qui ne doivent pas être comptées ni comme prédicats ni comme mots. Les parenthèses ( ) servent à entourer les mots qui doivent être intégrés dans le compte du nombre total de mots mais ne doivent pas être marqués comme prédicats. Entrent dans ce cas les énoncés inachevés, les énoncés et/ou mots inaudibles, les marqueurs discursifs (e.g. « vous voyez »), les interjections (e.g. « bon ») qui ne doivent pas être traitées comme des adjectifs (donc des prédicats) mais qui sont considérées comme des mots contrairement aux pauses remplies non-lexicales (e.g. « bah », « hein », etc.), et les noms propres pour éviter les problèmes d'étiquetage morphosyntaxique. Par exemple, voici le prétraitement de la phrase analysée plus haut :

> *le plus beau jour de ma vie (bon) bah (réfléchissons) c'est le jour de mon mariage (voilà) il y a 52 ans bientôt (donc voilà) bien que ça a été un mariage tout à fait simple parce que [je ne] je n'avais plus mes parents donc quand on s'est mariés on était 12 personnes (donc vous voyez)*

Ce corpus prétraité manuellement a été soumis au calcul automatique de la DI avec la version 1.2 de *Densidées*. Deux experts ont travaillé individuellement chaque texte en notant prédicat et argument de chaque mot sur Excel, un troisième examinateur a vérifié leur analyse. Le résultat obtenu par *Densidées* est donc vérifié par ces trois experts, pour mesurer le coefficient de corrélation de l'analyse manuelle et automatique.

3 Résultats

La figure 3 montre les résultats obtenus. On peut noter, entre la densité des idées calculée manuellement et automatiquement, pour les 40 textes, un coefficient de corrélation de 0.972, là où CPIDR obtenait 0.942.

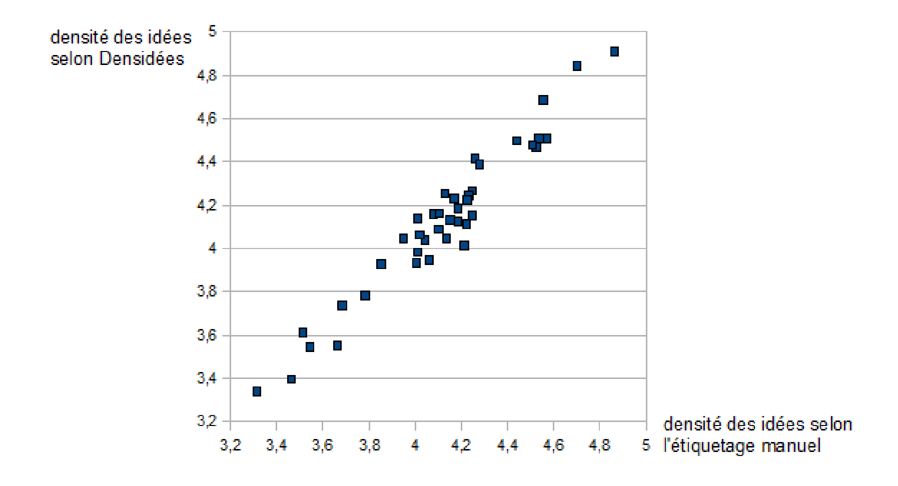


Figure 3. Représentation de la densité des idées calculée automatiquement en fonction de la densité des idées calculée manuellement pour chacun des 40 textes du corpus

Pour une évaluation plus fine du logiciel, nous avons choisi de déterminer le taux de faux négatifs (prédicats non étiquetés comme tels) et de faux positifs (non-prédicats étiquetés comme prédicats) : respectivement 2.7% et 3.1%. Comme la formule de densité des idées fait intervenir le nombre total de prédicats, ces deux types d'erreurs se compensent, pour arriver à un taux d'erreur moyen de 0.5% sur le nombre de prédicats. Le corpus a alors été séparé en une base de test (correspondant à 10 sujets pour assurer une variété dans les scores de DI) de 3728 mots et 1548 prédicats et une base de validation de 10211 mots et 4199 prédicats. La base de test a été utilisée pour évaluer la pertinence de chaque règle, en testant l'effet de sa suppression. Pour évaluer la qualité d'un étiquetage automatique, on calcule la F-mesure, qui se base sur la précision (i.e. proportion de prédicats corrects parmi les prédicats trouvés automatiquement) et le rappel (i.e. proportion des prédicats corrects trouvés par *Densidées* sur l'ensemble des prédicats corrects). En n'utilisant que la règle 200, qui étiquette les conjonctions, numéraux, déterminants, prépositions, adjectifs, adverbes et verbes comme prédicats, on obtient 29 faux négatifs et 1003 faux positifs, ce qui correspond à une F-mesure de 0.747 sur la base de test. Si l'on prend en compte l'ensemble des 35 règles de la version 1.2, on arrive à une F-mesure de 0,975.

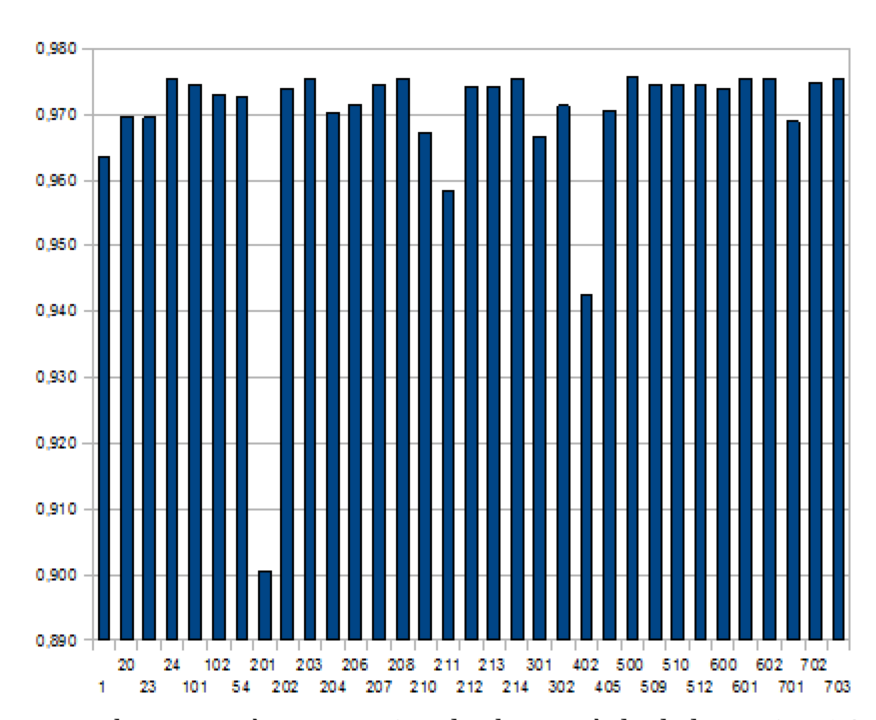


Figure 4. F-mesure obtenue après suppression de chaque règle de la version 1.2 de *Densidées*

La figure 4 illustre la dégradation de la F-mesure induite par la suppression de chaque règle. Ainsi, on constate par exemple qu'en retirant la règle 201 dédiée à l'étiquetage des déterminants "un", "une", "le", "la", etc. (qui ne sont pas des prédicats), la F-mesure décroît à 0.900. Inversement, la suppression de certaines règles (24, 203, 208, 214, 601, 602, 703) n'a aucun effet, voire améliore légèrement la F-mesure en supprimant un faux positif pour la règle 500.

Même si la suppression de ces règles semble n'avoir aucun effet sur le corpus de test, elle pourrait en avoir sur d'autres textes, ce qui explique que nous les laissons dans *Densidées*. En effet, certaines de ces règles sont plus adaptées pour des discours écrits (comme la 214 qui considère « si ... alors » comme un seul prédicat). D'autres, à l'inverse, sont prévues pour le discours oral, mais ciblées sur des marqueurs discursifs particuliers que les locuteurs n'utilisent pas nécessairement : la règle 602 considère par exemple que « donc » n'est pas un prédicat après le verbe « dire ».

Munis de ce score de qualité qu'est la F-mesure, nous proposons la méthodologie suivante pour l'ajout de nouvelles règles ou la modification de règles existantes dans *Densidées*. Nous utilisons la base de test à titre d'exploration pour évaluer l'évolution de la F-mesure suite à des modifications du programme, la base de validation sert quant à elle à valider la pertinence des modifications, en vérifiant que les modifications proposées à partir du corpus de test ne sont pas biaisées par les spécificités linguistiques de ce corpus. Par exemple, par rapport à la version 1.2, parmi toutes les modifications de règles testées, une seule (modification de la règle 301), raisonnable du point de vue linguistique, a permis d'améliorer la F-mesure en atteignant 0.978 sur le corpus de test. Sur la base de validation, cette modification a permis de passer d'une F-mesure de 0.969 à 0.972. Ainsi, elle sera intégrée dans la version 1.3 de *Densidées*.

4 Discussion

Un entretien oral de 300 mots nécessite environ 25 minutes de transcription, 10 minutes de parenthésage et 35 minutes d'étiquetage manuel des idées. Ainsi, *Densidées* permet de diviser par deux le temps total nécessaire à l'évaluation manuelle de la densité des idées. Il permet surtout de normaliser l'étiquetage en évitant les spécificités d'étiquetage des experts humains sur certains mots difficiles à étiqueter. Précisons que le temps de parenthésage est principalement dû à la relecture attentive du texte nécessaire à la détection des passages vides de sens ou correspondant à des idées répétées. Rappelons que notre corpus n’était pas créé pour l’étude du calcul automatique de la densité des idées. Si l’objectif est bien déterminé au départ (e.g. application médicale) et que les entretiens sont réalisés dans l'optique exclusive de calculer la densité des idées, la transcription obéit à des contraintes ciblées, et peut se faire plus rapidement en conjonction avec l'étiquetage (e.g. les passages incompréhensibles placés entre crochets ne sont alors simplement pas transcrits).

Il faut aussi noter, à propos des bonnes performances obtenues par *Densidées*, que les règles d’étiquetage de mots en prédicats ou non-prédicats présentent une certaine robustesse par rapport à d’éventuelles erreurs au cours de l’étiquetage morphosyntaxique. En effet, une erreur d’étiquetage d’un verbe pris pour un adjectif n’aura souvent aucune conséquence sur l’étiquetage réalisé par *Densidées*, puisque verbes et adjectifs sont généralement tous deux considérés comme des prédicats (par la règle 200).

Si la densité des idées est une méthode efficace pour mesurer la quantité d’informations dans un discours, de nombreuses applications de l’analyse qualitative fine qu’offre l’analyse prédicative sont délaissées, du fait que les arguments ne sont pas pris en compte dans cette méthode. Par exemple, le calcul du *décalage* (i.e. produit par le partage du même argument par différentes propositions, marquant la cohésion du discours et sa complexité sémantique) (Duong et *al.*, 2000). Aussi, le calcul du *prédicat de premier rang* (i.e. qui n’implique que des arguments objets) et du *prédicat de rang supérieur* (i.e. implique également ou exclusivement des arguments propositionnels). Étant donné que le prédicat de rang supérieur a un coût cognitif plus important, ce type de calcul serait utile pour l’étude psycholinguistique. Pour contourner ces limites, une méthode probabilistique peut être envisagée, en mesurant par exemple le nombre et le type d’arguments que peut avoir un prédicat (e.g. verbe transitif « vendre » comporte 3 places d’argument : agent, objet, récepteur). *Densidées* version 1.2 fait ses premiers pas vers une interprétation qualitative du résultat de la DI, en offrant les résultats détaillés des règles utilisées pour le calcul (e.g. le taux important de l’utilisation de la règle 211 refléterait un discours construit principalement autour de la négation, etc.).

5 Conclusion

Le logiciel *Densidées* fournit actuellement une approximation tout à fait satisfaisante de la densité des idées d'un discours oral transcrit selon la méthodologie que nous proposons ici. Cette méthodologie fait intervenir une détection humaine des idées répétées, et il pourrait être envisagé d'aborder ce problème de

façon automatique. Toutefois, nous pensons que l'effort d'étiquetage humain des idées répétées pendant la transcription constitue un effort minime, et envisageons plutôt pour de prochaines versions du logiciel d'améliorer l'étiquetage morphosyntaxique sur lequel se basent les règles de *Densidées*, en faisant appel au logiciel Cordial au lieu de TreeTagger.

On peut également noter que l'étiquetage automatique des prédicats dépend fortement des habitudes langagières récurrentes des locuteurs (ajout de « quoi » en fin de phrase par exemple). Ainsi, compléter cette approche par règles avec une partie statistique (par exemple pour détecter des mots fréquents inattendus) pourrait aider à repérer ces habitudes langagières, et proposer automatiquement de nouvelles règles adaptées.

Enfin, la densité des idées a d'autres applications citées plus haut, dont la mesure du niveau de technicité d'articles scientifiques en langue anglaise. Il serait intéressant de calculer la densité des idées de corpus écrits en français pour tenter d'identifier certains genres de textes associés à une densité des idées très élevée ou au contraire très basse. La densité des idées pourrait aussi s'ajouter aux paramètres pertinents pour choisir des textes en fonction de leur niveau de technicité en enseignement du français langue étrangère (Thomas, 2009).

Références

ANDERSON J. (1976). *Language, memory and thought*. Hillsdale, NJ: Erlbaum Associates.

BLANCHE-BENVENISTE C. (1998). *Approches de la langue parlée en français*. Paris : Ophrys.

BROWN C., SNODGRASS T., KEMPER S., HERMAN R., COVINGTON M. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540-545.

CHAND V., BAYNES K., BONNICI L., TOMASZEWSKI FARIAS S. (2010). Analysis of idea density (AID) : A manual. University of California at Davis.

COIRIER P., GAONAC'H D., PASSERAULT J.-M. (1996). *Psycholinguistique textuelle* : approche cognitive de la compréhension et de la production des textes. Paris: Armand Colin.

COVINGTON M. (2009). Idea Density: A potentially informative characteristic of retrieved documents. *Proceedings, IEEE SoutheastCon*.

COVINGTON M., RIEDEL W., BROWN C., HE C., MORRIS E., WEINSTEIN S., et *al.* (2007). Does ketamine mimic aspects of schizophrenic speech? *Journal of Psychopharmacology*, 21, 338-346.

DENHIÈRE G. (1984). *Il était une fois... Compréhension et souvenir de récits*. Lille: Presses Universitaires de Lille.

DUONG A., SKA B., POISSANT A., JOANETTE Y. (2000). Effet du vieillissement de la scolarité et du stimulus sur la production de narrations. In *Le vieillissement cognitif normal. Vers un modèle explicatif du vieillissement*, 137-154. Bruxelles : Edition de De Boeck Université.

FILLMORE C. (1968). The case for case. In *Universals in linguistic theory*, 1-88. New York: Holt, Rinehart, and Winston.

FRANÇOIS J. (1991). Pertinence linguistique des représentations propositionnelles de la sémantique cognitive. *Sémiotiques*, 1(1), 69-80.

FRANÇOIS T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. *Actes de RECITAL 2009*.

FREGE G. (1967). *The basic laws of arithmetic*. Berkeley: University of California.

FREGE G. (1971). *Écrits logiques et philosophiques*. Paris : Presses Universitaires de France.

GHIGLIONE R. (1982). Analyse propositionnelle et modèles argumentatifs. *Connexions*, 38, 89-106.

GHIGLIONE R., KEKENBOSCH C., LANDRÉ A. (1995). *L'analyse cognitivo-discursive*. Grenoble: Presses Universitaires de Grenoble.

GOODGLASS H., KAPLAN E. (1983) *The assessment of aphasia and related disorders*. Philadelphia : Lea and Febiger.

JAKOBSON R. (1963). *Essais de linguistique générale*. Paris : Éditions de Minuit.

KEMPER S., GREINER L., MARQUIS J., PRENEVOST K., MITZNER T. (2001). Language decline across life span: findings from the Nun study. *Psychology and Aging*, 16(2), 227-239.

KINTSCH W. (1974). *The representation of meaning in memory*. Hillsdale, NJ : Erlbeaum.

KINTSCH W., KEENAN J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3), 257-274.

KINTSCH W., KEENAN J. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

LE NY J.-F. (1979). La sémantique psychologique. Paris : Presses Universitaires de France.

LE NY J.-F. (1987). Sémantique psychologique. In *Problèmes de psycholinguistique*, 13-42. Bruxelles : Pierre Mardaga.

LE NY J.-F. (1989). *Science cognitive et compréhension du langage*. Paris : Presses Universitaires de France.

LEE H., BARKAT-DEFRADAS M. (2009). La densité des idées : un modèle d'analyse du discours pertinent pour le diagnostic précoce de la maladie d'Alzheimer ? *Actes des 8ème Rencontres Jeunes Chercheurs en Parole*.

LIU H. (2004). MontyLingua : An end-to-end natural language processor with common sense. Available at: web.media.mit.edu/~hugo/montylingua

MEURS M., DUVERT F., BÉCHET F., LEFÈVRE F., DE MORI R. (2008). Annotation en Frames Sémantiques du corpus de dialogue MEDIA. *Actes de TALN 2008.*

MOREAUD O., DAVID D., CHARNALLET A., PELLAT J. (2001). Are semantic errors actually semantic ? Evidence from Alzheimer's disease. *Brain and language*, 77, 176-186.

SCHMID H. (1994). Probabilistic Part-of-Speech tagging using decision trees. In *New Methods in Language Processing*, 154-164. London : UCL Press.

SNOWDON D., KEMPER S., MORTIMER J., GREINER L., WEKSTEIN D., MAYKESBERY W. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the Nun Study, *JAMA*, 275, 528-532.

TAKAO A., PROTHERO W., KELLY G. (2002). Applying argumentation analysis to assess the quality of university oceanography students' scientific writing. *Journal of Geoscience Education*, 50, 40-48.

THORSON E., SNYDER R. (1984). Viewer recall of television commercials: structure of commercial scripts. *Psychological Review*, 85, 363-394.

TURNER A., GREENE E. (1977). The construction and use of a propositional text base. *Technical report*, 63. Boulder: Institute for the study of intellectual behavior, University of Colorado.